MDDT Summary of Evidence and Basis of Qualification (SEBQ)

MDDT SUMMARY OF EVIDENCE AND BASIS OF QUALIFICATION DECISION FOR
RUBRIC FOR APPLYING CVSS TO MEDICAL DEVICES
VERSION: 0.12.04 – SEPTEMBER 3, 2019

BACKGROUND

MDDT NAME*:* RUBRIC FOR APPLYING CVSS TO MEDICAL DEVICES

SUBMISSION NUMBER*:* Q171974

DATE OF SUBMISSION*:* 11/14/2017

CONTACT: Ms. Penny Chase
The MITRE Corporation
202 Burlington Rd
Bedford, MA 01730
UNITED STATES

TOOL DESCRIPTION AND PRINCIPLE OF OPERATION

The rubric, version 0.12.04, is structured as a series of questions at various decision points. Each portion of the CVSS vector has its own rubric and series of structured questions. Each answer should be recorded by the analyst. Many answers provide direct suggestions for how to fill out a portion of the CVSS vector; typically, the analyst is expected to use the first vector suggestion that is associated with the question(s), as the questions are organized in a way that prioritizes answers with the most significant contribution to the CVSS score. Other questions ask for additional information that does not directly affect the CVSS vector, but the answers could be used by the manufacturer/HDO or other stakeholder in conducting additional risk analysis. By design, the rubric can cause the analyst to "skip" some subsequent questions that become irrelevant when the analyst follows a different branch. The rubric also allows the analyst to record when an answer is unknown; the worst-case metric value is then used for the scoring engine.

Finally, when the answer to a question suggests that the vulnerability might have an adverse effect on patient safety, there is an explicit notice that the analyst might need to perform a safety-oriented hazards analysis to determine whether the issue must be reported to FDA/CDRH as covered in the Post-Market Guidance. Such items are marked as PIPS, an informal acronym that stands for "Potential Impact to Patient Safety."

In addition to the series of structured questions, each portion of the CVSS vector has a Decision Flow diagram and an Extended Vector table. The Decision Flow diagram depicts

the decision flow logic of the series of structured questions in a graphical format. The Extended Vector table specifies the extended vector that results from answering the series of structured questions: the table defines the corresponding extended vector element and its allowed values for each question.

For better results, the scoring exercise should involve consultation with a group of subject matter experts (SMEs), not just a single analyst. From the perspective of patient safety, at a minimum, the following knowledge areas should be shared across the entire group, although it is expected that each SME might only be an expert in one area:

- Cybersecurity and privacy
- Device engineering, design, and architecture
- Patient health impact from resulting hazards
- HDO device usage scenarios and clinical workflow impact
- Information technology integration and interoperability

Once the analyst applies the rubric to a particular vulnerability or security concern for a medical device, the following information could be provided as output:

- CVSS score (between 0 and 10.0), as calculated using the FIRST CVSS v3.0 specification;
- CVSS vector (a set of tuples), as defined in the FIRST CVSS v3.0 specification;
- Answers to the rubric's related questions, which may help guide or understand healthcare-specific considerations for the larger risk analysis. Currently, these are being represented in a way that allows creation of an "extended vector" that has the same syntax as a CVSS vector; each measure's code begins with "X." An example scorecard is included in the rubric.

QUALIFIED CONTEXT OF USE

The Mitre "Rubric For Applying CVSS To Medical Devices - Version: 0.12.04 – September 3, 2019" is qualified for the evaluation and justification of patient-centric, situational impact and urgency characteristics in time-sensitive postmarket vulnerability disclosures of medical devices, when supporting the FIRST CVSS V3.0 standard. The accompanying vector string should always be published together with the score for any such evaluation.

SUMMARY OF EVIDENCE TO SUPPORT QUALIFICATION

The sponsor conducted two series of pilots to gather evidence on performance characteristics of the CVSS supplemental rubric, the first prior to the MDDT proposal submission with two infusion pump manufacturers, and the second after receiving feedback from FDA on the proposal (mainly questions about the applicability of the rubric to a broader range of devices).

For the first pilot, the sponsor identified two unnamed infusion pump manufacturers who use CVSS as part of their medical device development lifecycle. The goal was to assess the differences between using their existing cybersecurity vulnerability assessment processes against the process of using the rubric and CVSS V3.0 together.

The sponsor presented each manufacturer with two vulnerabilities that had been discovered in their own devices, and then presented one or two "theoretical vulnerabilities" – based on publicly disclosed vulnerabilities in the same or similar type of medical device. The manufacturers assessed these vulnerabilities using their current process of CVSS scoring (one uses CVSS version 3 and the other uses version 2.

The sponsor then provided the manufacturers with the rubric and they rescored the same vulnerabilities. After each vulnerability was scored using the rubric, the sponsor used a questionnaire to qualitatively assess the differences between the outcomes of the processes. The data gathered were the extended vector, CVSS vector assignments, and CVSS score for the Base Metric Group, as well as the questionnaire results and qualitative assessments based on observing the scoring process and discussions with the participants.

For the second pilot, the sponsor wanted to demonstrate that the context of use for the CVSS supplemental rubric is broader than infusion pumps and gather additional evidence on performance characteristics, including consistency of scoring, validity of the rubric versus expert assessment, and usability of the rubric.

The sponsor identified two additional unnamed manufacturers who currently use CVSS as part of their security risk assessments. One manufacturer had four product divisions who used CVSS in different ways: the standard CVSS v3, a modified version of CVSS v2, and one had recently started to use the CVSS supplemental rubric. The product areas covered by the second pilot were: insulin pumps, radiological imaging devices, implantable cardiovascular devices, patient programmers for neuro-stimulators, and dialysis devices. The fairly wide spectrum of devices included categories which reflect the wide variety of user and operator viewpoints i.e personally worn devices, hospital infrastructure, implants, marketed product and specialized programmers used in physician offices.

For these manufacturers, the sponsor identified a few representative real-world vulnerabilities that had been discovered in their devices and previously scored with CVSS. These vulnerabilities were rescored using the rubric together with the CVSS v3.0 and the sponsor then compared the scores and CVSS vectors, discussing the differences guided by a questionnaire similar to the one used in the first pilot. The sponsor also provided one or two additional theoretical vulnerabilities that were scored by multiple teams at each manufacturer to assess consistency in scoring when using the rubric. The data gathered were the extended vector, CVSS vector assignments, and CVSS score for all metric groups[13], as well as qualitative assessments based on observing the scoring process and discussions with the participants.

Reference [13] In some cases computing the temporal score was skipped because in the scenario the vulnerability was newly discovered by the manufacturer so the temporal vector elements would have biased the score (e.g., there wouldn't be any type of remediation at the time a vulnerability is initially discovered).

Strength of Evidence

During the study the sponsor gathered evidence on using the CVSS supplemental rubric to score seven actual vulnerabilities and seven theoretical vulnerabilities. Although that's a small sample, FDA should put it in perspective: between October 2013 and August 2019, there were 68 ICS-CERT advisories for 152 vulnerabilities in medical devices. The Table 1 below provides summary data on the advisories and vulnerabilities. The table is divided into two periods showing an increase in vulnerability reporting following the publication of the Post- market Guidance on Dec. 28th 2016.

See Table 1 and Figure 1 below.

Time Frame	10/23/2013 – 12/28/2016	12/29/2016 – 3/31/2019
Number of Advisories	12	51
Total vulnerabilities disclosed in advisories	37	109
Average vulnerabilities per month	0.95	4.19
Companies	6	24
Mean vulnerabilities CVSS scores	7.30	6.87

Table 1 – Summary data on advisories and vulnerabilities

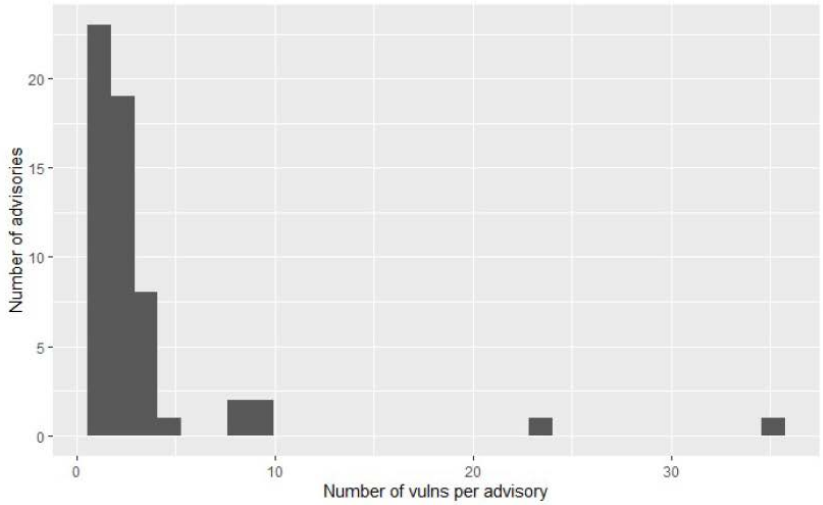

Figure 1 – Distribution of number of vulnerabilities per advisory (Source: CVSS v3 Data)

These data suggest the sponsor was dealing with a relatively small population of disclosed vulnerabilities.

The evidence gathered focused on the following performance characteristics of the rubric:

- Does the rubric scoring produce results that were expected? When using the rubric to score actual vulnerabilities, did the rubric produce results that are similar to the

original assessment, or if different, did the rubric produce a result that more accurately reflected the severity of the vulnerability?

- Does the rubric scoring produce more consistent results than CVSS scoring without the rubric, i.e., if multiple teams with similar product awareness within a single manufacturer independently scored the vulnerability, were the results similar? How does this compare to the baseline of comparing ICS-CERT advisory and NVD CVSS scores for medical device vulnerabilities?
- Does the rubric enable more efficient scoring, i.e., did participants feel that the rubric aided in reaching consensus more quickly?
- Does the rubric enable more effective communication of the potential patient safety impact of a vulnerability?

The sponsor discussed each of these performance characteristics in turn. For the first two, characteristics the sponsor compared CVSS scores and vectors. Although CVSS scores look like continuous data, the CVSS algorithm takes ordinal data (the vector values) and "assigns relative importance rankings as ratio values." This means that everyone should keep in mind that users of this rubric are really working with qualitative data in the use of this tool.

The sponsor looked at the number of identical vector elements for the NVD and ICS-CERT vector assignments and found that roughly 37% of the vectors are identical and about 26% differ in a single element, and about 40% differ in two or more elements. See Figure 2 below.

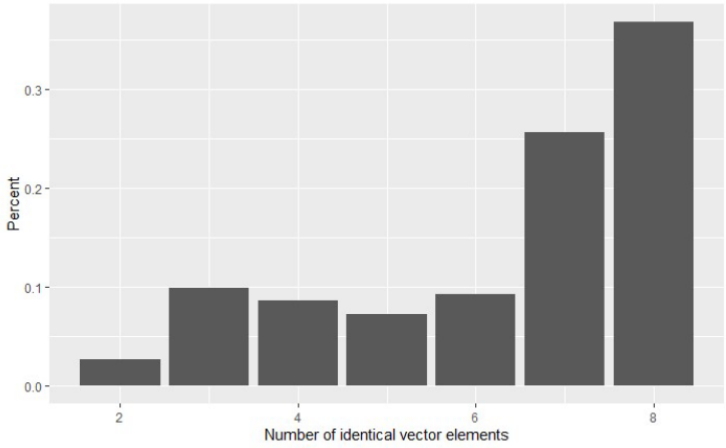

Figure 2 – Number of differences in vector elements (NVD vs ICS-CERT)

The following table shows the agreement between NVD and ICS-CERT vector assignment by vector element:

Vector element	Percent agreement
AV	82%
AC	76%
PR	80%
UI	95%
S	91%
C	66%
I	72%
A	74%

Table 2 – Percent agreement in vector element assignment (NVD vs ICS-CERT)

It can be seen from Table 2 that the greatest differences in vector assignment were observed in the confidentiality, integrity, and availability impacts.

Does the rubric scoring produce expected results?

During both pilots the sponsor worked with the manufacturers to identify vulnerabilities in their products that had been previously disclosed (or identified during design/development) and scored with CVSS. The sponsor asked the manufacturers to re-score the vulnerabilities with the rubric in order to compare the results (see Table 3 below).

Original Vector	Original Score	Original Ranking	Rubric Vector	Rubric Score	Rubric Ranking
AV:P/AC:H/PR:N/UI:N/S:C/C:H/I:N/A:N	4.9	Medium	AV:P/AC:L/PR:N/UI:N/S:U/C:L/I:N/A:N	2.4	Low
AV:N/AC:L/PR:L/UI:N/S:C/C:H/I:H/A:H	9.9	Critical	AV:N/AC:L/PR:L/UI:N/S:C/C:H/I:H/A:H	9.9	Critical
AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:L/A:L	9.9	Critical	AV:N/AC:L/PR:L/UI:N/S:C/C:H/I:H/A:H	10	Critical
AV:N/AC:L/PR:N/UI:R/S:U/C:H/I:H/A:H	8.8	High	AV:N/AC:L/PR:L/UI:R/S:U/C:L/I:H/A:H	7.6	High
AV:A/AC:H/PR:N/UI:N/S:U/C:N/I:H/A:N	5.3	Medium	AV:A/AC:L/PR:N/UI:R/S:U/C:H/I:H/A:H	8	High
AV:P/AC:L/PR:N/UI:N/S:C/C:H/I:H/A:H	7.6	High	AV:P/AC:L/PR:N/UI:N/S:U/C:H/I:H/A:H	6.8	Medium

Table 3 – Rescoring previously disclosed vulnerabilities

50% of these vulnerabilities received the same qualitative ranking during the re-scoring. In these cases the vector assignments were identical or differed in one of the CIA impacts. In two of the cases when the ranking differed, the re-scored ranking was lower because the rubric caused the participants to think differently about the scope element. In the other case the re-scored ranking was higher because the rubric forced the participants to systematically think through the CIA impacts and identify impacts on data that they hadn't considered in the original assessment because the manufacturer's team focused on the researcher's claims and did not consider other potential technical impacts which they alone knew about.

During the first pilot, the sponsor also provided theoretical vulnerabilities that were first scored with CVSS without the rubric and then with the rubric. In all cases the qualitative rankings were identical. In one case the scores were identical, but the participants felt there was more consensus in reaching the score when using the rubric. When the vector assignments differed, the participants felt that the rubric provided greater clarity and a more systematic approach in making the assignment (e.g., distinguishing between cases in the Attack Vector element).

Given the above discussion on the differences between NVD and ICS-CERT scoring, it's not surprising that the CVSS and rubric plus CVSS vector assignments and scores differed; what is significant is that in all cases the participants felt that the rubric assignments/score more accurately reflected the severity of the vulnerability in medical devices.

Does the rubric scoring produce more consistent results?

During the second pilot the manufacturer teams were large enough to divide into groups to independently score the same vulnerability (either an actual vulnerability or a theoretical vulnerability) to assess whether the rubric promoted more consistent vector assignment and scoring (see Table 4 below).

Team 1 Vector	Team 1 Score	Team 1 Ranking	Team 2 Vector	Team 2 Score	Team 2 Ranking
AV:N/AC:L/PR:L/UI:N/S:C/C:L/I:N/I:N	5	Medium	AV:N/AC:L/PR:N/UI:N/S:C/C:L/I:N/I:N	5.8	Medium
AV:N/AC:H/PR:N/UI:R/S:C/C:L/I:H/A:H	8.2	High	AV:N/AC:H/PR:N/UI:R/S:C/C:N/I:H/A:H	8	High
AV:N/AC:H/PR:N/UI:R/S:U/C:H/I:H/A:H	7.5	High	AV:N/AC:H/PR:N/UI:R/S:U/C:H/I:H/A:H	8.3	High

Table 4 – Multiple scoring of vulnerabilities

In all cases the CVSS Qualitative rankings were the same and the vector assignments differed by one element, demonstrating a good level of consistency compared with the baseline comparison between NVD and ICS-CERT. Even more significant is that in the discussions after the scoring, as the teams compared how they used the rubric, they reached amicable resolution on which team had the correct interpretation.

Does the rubric enable more efficient scoring?

After the vector assignment and scoring exercises during the pilots the sponsor discussed the process using the questionnaire in Appendix B of the qualification package to frame the conversation. In addition to assessing the scoring itself, the sponsor wanted to understand if the participants felt that the rubric made the vulnerability assessment easier or more complex. The participants observed that the rubric was more complex and detailed than CVSS alone, but that it refined the discussions, forced the teams to think systematically, and made the scoring process more repeatable and consistent. Some issues arose during the second pilot that led to some minor modifications to the rubric. The version of the tool shown reflects the modified approach.

In assessing the technical impacts, the rubric asks a series of questions about the impact on different data/functionalities for confidentiality, integrity, and availability in turn. During the scoring exercises the sponsor observed that the this was both repetitious and that when discussing the impact on one member of the CIA triad, other members were also unavoidably discussed. It was suggested during one of the pilot sessions to restructure the rubric to collapse the CIA technical impacts into a single decision tree.

The other part of the rubric (and CVSS itself) that presented challenges in vector assignment was the “Change of Scope” element. The FIRST CVSS SIG has had discussions about the scope element in their forum and the sponsor leveraged this discussion to provide additional considerations for the scope element in the rubric.

Does the rubric enable more effective communication of the potential patient safety impact of a vulnerability?

In the discussions of the rubric assessments during the pilots, the manufacturer teams observed that the rubric contributed to more effective conversations between the team members during the scoring. One of the manufacturers observed that the rubric would be useful in having conversations with management: a tool produced by an independent third party would help have these conversations focused on security engineering.

In addition to internal communications, the pilot participants believed that the rubric could drive communications with their customers. They observed that rubric questions and extended vector could “prove” how they arrived at the score and shows that a broad range of issues were considered during the assessment.

During the second pilot the sponsor included the environmental metric group in the vector assignments/scoring and the manufacturers observed that it would be useful in communicating what a hospital could do (compensating controls) and help identify ‘low hanging fruit’ options to reduce the severity and potential impact of a vulnerability. Indeed, after the pilot session, one manufacturer has started to incorporate the rubric in product cybersecurity white papers for their customers.

Summary

During these pilot studies with four manufacturers the sponsor gathered evidence on the performance of the CVSS supplemental rubric. That evidence suggests that using the supplemental rubric to assess the severity of vulnerabilities in medical devices aligns with subject matter experts’ (SMEs) assessment of the vulnerabilities. Some of the evidence compared the assessments of previously analyzed vulnerabilities with the assessment by the rubric. In half the cases there was substantial agreement between the two assessments and in the remainder, the SMEs, who had previously conducted the thorough analysis as part of the disclosure process, believed that the scoring with the rubric produced an assessment that more accurately reflected the severity of the vulnerability.

During two of the pilot studies the sponsor was able to gather evidence on consistency through multiple teams assessing the same vulnerabilities, which demonstrated that the supplemental rubric contributed to greater consistency in CVSS vector assignments and scoring. This approach showed that for a given knowledge base of the affected platform the rubric allowed convergence to a more accurate consensus score. Even when there was initial disagreement, in subsequent discussions the rubric enabled the teams of SMEs to quickly reach consensus on which one of the different vector assignments was correct.

The manufacturers' teams believed that the supplemental rubric helped them achieve consensus during vulnerability discussions. The rubric provided a systematic approach to assessing the vulnerabilities. Even if the rubric was perceived as more complicated and detailed than a CVSS assessment without the rubric, the pilot participants felt that the rubric helped them come to agreements more quickly and that discussions were focused on healthcare and patient safety impacts.

Finally, the manufacturer teams felt that the rubric's real value was in fostering consistent communications. The rubric provides more than a number, the score: it documents the scoring process and reflects the clinical end-user environment in assessing exploitability and technical impacts.

In addition, by using the environmental metric group in the vulnerability assessments, the manufacturers found the rubric to be a useful tool for communicating to their customers the value of using recommended mitigations (compensating controls) to address the potential impacts of a vulnerability.

DISCUSSION OF THE EVIDENCE STRENGTH TO SUPPORT QUALIFICATION

The rubric together with the FIRST CVSS V3.0 standard seems to allow iteration to a more accurate estimate of the patient risk aspects of a cyber vulnerability between similarly knowledgeable SME's. It also permits a less knowledgeable third party to quickly see how the more knowledgeable SME's arrives at their conclusion and the data show the ensuing discussions are more productive and more likely to reflect the instantaneous risk profile of the situation. Accuracy and precision of the CVSS V3.0 are improved through the application of the rubric.

The ability of different teams within a single manufacture to iterate to convergence on scoring provides FDA with some insight into the potential predictive accuracy of the outcomes when used by different manufacturers on their own devices to reach "amicable' agreement among stakeholders. This indicates a reasonably good predictive nature for the rubric.

ASSESSMENT OF ADVANTAGES/DISADVANTAGES OF QUALIFICATION

There will remain some variability in the way a vulnerability's impact is perceived since each stakeholder has a different 'loss' calculus, risk appetite and risk management process.

The data from the sponsor's studies show that the use of the rubric together with the CVSS V3.0 standard by multiple stakeholders may permit a more transparent score and communication of impact and urgency, reflecting the instantaneous exploitability, of a single vulnerability. The additional use of the vector string underwriting that score may well permit a more rapid 'meeting of the minds' of the community of affected parties by allowing highly targeted questions to be posed between them reflecting the 'local' risk 'landscape' evaluation by each party. This would be a great advantage over the somewhat arbitrary results achieved by use of the CVSS V3.0 alone. Additionally, this will allow a more consistent application of FDA's post market policy by manufacturers, resulting in fewer needed interventions by FDA in these matters. This was the FDA's intention in drafting the post market guide in 2016 and this rubric may well operationalize that guidance. Additionally the use of the rubric may well facilitate an industry wide consensus-management culture of vulnerability disclosures in which all stakeholders can learn to treat the process as just another "…opportunity for improvement". Reducing the variability of risk assessment and risk evaluation creates trust and transparency, allowing a lighter touch from regulators. Only when fundamental disagreements emerge between stakeholders, or in cases of dire public health circumstances, will FDA need to be fully engaged and FDA can quickly re-engage, if needed, when this semantic framework is part of the industry culture.

The data are still relatively 'thin' for the rubric's use but inside FDA we see a consensus emerging as trusted relationships are increasing the predictability of outcomes based on stable risk estimation techniques such as this. FDA is grateful for the efforts expended by participant manufacturers, whoever they were, in the sponsor's studies and FDA understands well the sensitive liability and regulatory concerns and reputational concern for participation in these pilots. As FDA's experience with the use of this rubric in the post market increases, we may acquire our own data moving forward in a quasi-Bayesian approach.

This technique is not suitable for estimating the impact and urgency of a 'chained' vulnerability attack, where a series of individual vulnerabilities in a single platform are used to systematically degrade a security architecture by simultaneously lowering the architecture's defenses until the attacker has revealed a hitherto invisible attack route. This type of very sophisticated attack on a security architecture will require a different tool to estimate its adverse impact and urgency on the overall architecture. This would be something like estimating the score of system cyber-resilience to multiple coordinated single attacks. The data do not yet support this rubric as a useful tool for this purpose.

No well accepted additive 'algebra' exists yet for accreting risk estimation of vulnerabilities into measures of system resilience. Nor is there any consensus in security engineering for the "…single point of failure" design heuristic so commonly observed in safety engineering. Nevertheless, good quality security risk estimations of single vulnerabilities

may facilitate rapid response to them, thus closing off access to chained attacks routes. We should discourage regulatory reliance on this tool for all chained attack analyses for the time being.

This rubric is also appearing in FDA pre-market submissions where the threat-model provided is based on the traditional tabular approach used in traditional safety risk estimation i.e. listing the threat, the unwanted outcome, the estimation of the risk, the mitigation and the re-estimation of the risk after mitigation.

This seems to be a way for manufacturers to 'justify' in a submission the engineering approach taken for addressing a given theoretical cyber-weakness in their design. It attempts to contrast the net benefit of their approach to what would happen without their approach. In other words, it seems to assert a level of "…non-inferiority to doing nothing" as being something useful. This use-case of the rubric is not helpful for this type of justification and may distract reviewers from asking the right questions about security architecture sufficiency. This approach might bear fruit in the future as methods to assess system cyber-resilience emerge.

FDA should not qualify this use-case in the pre-market as a valid regulatory use of the rubric. Only when an architecture-based approach for estimating the cyber resilience of the defense-in-depth of the design should that be considered. We should discourage all pre-market regulatory reliance on the tool for the time being.

CONCLUSIONS

In balancing the risks and benefits of qualifying the tool as an MDDT as outlined in the Context Of Use statement and as illustrated in greater detail above, the benefits greatly outweigh the risks. That balance will likely increase toward more benefit moving forward. FDA should simultaneously recognize the FIRST CVSS V3.0 standard under its statutory authorities and qualify this rubric under its MDDT guidance nearly simultaneously.

CONTACT INFORMATION FOR ACCESS TO TOOL

The rubric may be obtained from the Mitre Corp. Website at :- https://www.mitre.org/sites/default/files/publications/pr-18-2208-CVSS-medical-device-rubric-v0.12.04.pdf

The CVSS V3.0 user guide and access to the calculator can be obtained from FIRST at:- https://www.first.org/cvss/v3.0/user-guide

End of summary